IM-Net for High Resolution Video Frame Interpolation: Supplementary Material

Tomer Peleg Pablo Szekely Doron Sabo Omry Sendik
Samsung Israel R&D Center
{tomer.peleg,pablo.sz,doron.sabo,omry.sendik}@samsung.com

S1. More Details about IM-Net's Architecture

In Table S1 we present a split of the computational cost between the various sub-networks. This split shows that a good balance is achieved between the different processing stages. Since the network aims at block-wise estimates, the decoders requires much fewer operations with respect to the encoders, thus contributing to the light-weight nature of the network.

	Siamese	Encoder	Decoder
Level 0	18.5%	40.42%	7.27%
Level 1	4.63%	7.94%	5.16%
Level 2	1.16%	1.98%	5.16%
Merging and Estimations	7.75%		

Table S1. Computational load of each part of the network.

S2. Running IM-Net on Original Vimeo Dataset

In Section 5.3 of the main body we mentioned that we ran our approach with simple pre-processing and post-processing steps on the original Vimeo dataset. We will now elaborate on these steps. First, we up-sampled the input frames by a factor of 3, using plain bicubic up-sampling. We then ran IM-Net as usual. As a post processing step, we first divided the values of the IMVF by 3, and also down-sampled the IMVF and occlusion map by a factor of 1.5. This set of steps yielded an effective block size of 4×4, instead of the usual 8×8.

S3. Additional Visual Results

In the main body of the paper we have shown single frame results, either from two in-house test sequences or from the of Vimeo triplet dataset. This allowed us to compare the level of artifacts such as halos, ghosts and break-ups, between different CNN-based video frame interpolation methods (VFI). To further demonstrate these differences we show more visual results from the Vimeo dataset in Section S3.1.

In the supplementary material we also include video results on two in-house test sequences. As IM-Net is intended for video processing applications, it is most meaningful to compare the video quality of frame rate up-converted sequences. When observing videos, temporal artifacts such as flicker and wobble tend to be more significant than blur or small halo introduced by inaccuracies in the VFI method. A description and analysis of the video results will be provided in Section S3.2.

S3.1. Results from the Vimeo Dataset

In Fig. S1 we show 8 examples for interpolated frames computed on the original (448×256 resolution) and super-resolved (1344×768 resolution) versions of the Vimeo dataset by IM-Net and two previous CNN-based methods – TOFLow and SepConv. This figure has a similar form as Fig. 4 in the main manuscript. From these examples we can see a large difference between the performance of the TOFlow and SepConv methods on the original low resolution input frames and on the super-resolved version of these inputs. When applying these methods on the super-resolved frames, they are prone to severe halo, ghost and break-up artifacts.

On the low resolution frames IM-Net is inferior to TOFlow and SepConv in terms of the introduced blur and halo. However, it keeps a similar frame interpolation quality for the super-resolved frames, hence significantly outperforming the two previous methods on high resolution scenes with strong enough motions. In the first example IM-Net preserves the shape of the bag. It avoids severe break-ups of the head in the second and fourth examples, and maintains fine geometrical structures in the fifth and sixth examples. IM-Net also avoids severe ghosts in the seventh example and keeps the shape of the gloved hand intact in the last example.

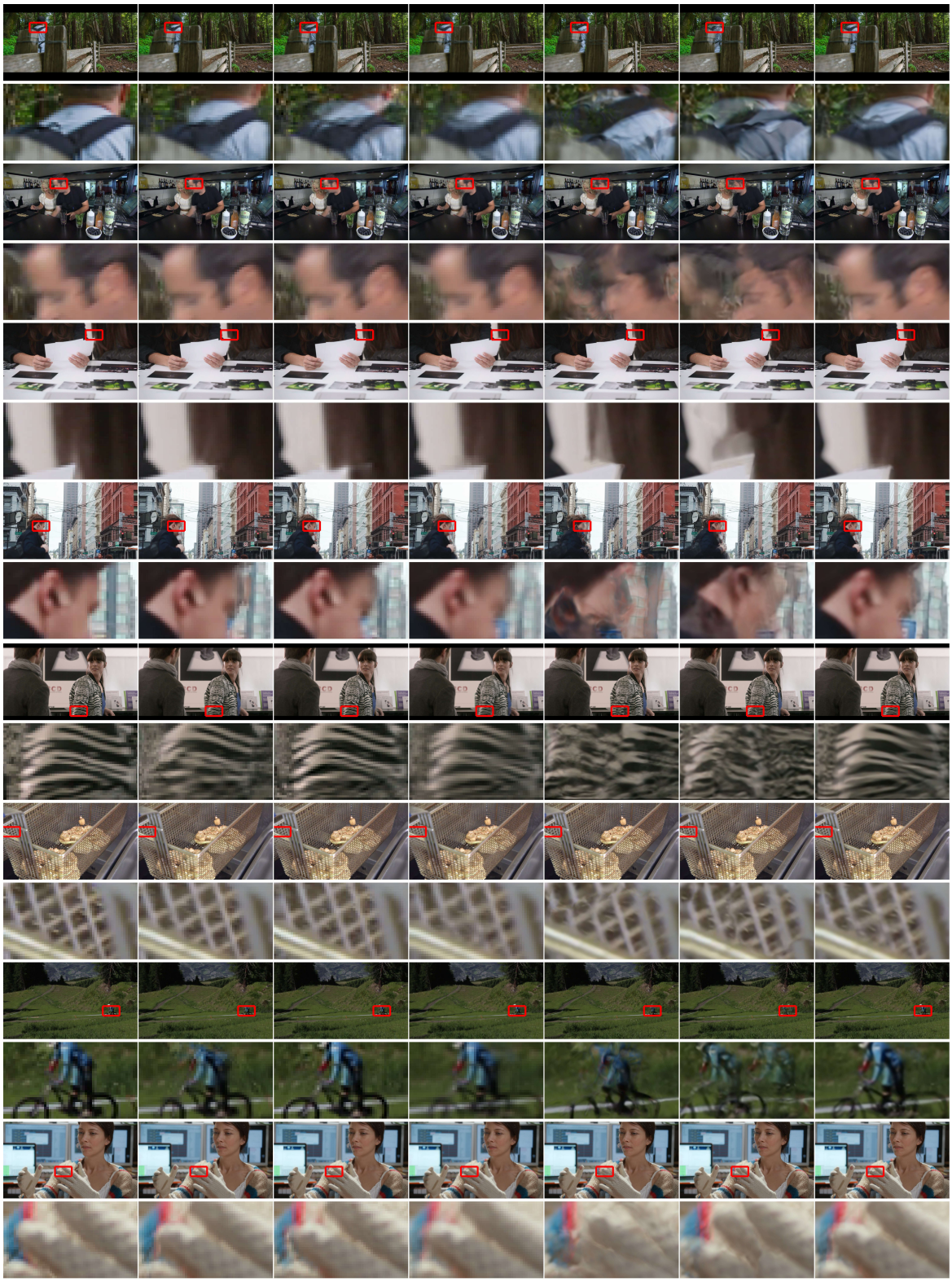

Figure S1. Example results from the Vimeo dataset (best viewed in color), from left to right: 448x256 ground-truth frame, interpolated frame synthesized by TOFlow on 448×256 inputs, SepConv on 448×256 inputs, IM-Net on 448×256 inputs, TOFlow on 1344×768 inputs, SepConv on 1344×768 inputs, and IM-Net on 1344×768 inputs. In each pair of rows we show the full frame on the top and zoom-in of a cropped interesting region (highlighted in a red box) on the bottom.

S3.2. Video Results

In the attached avi files[1] we show a side-by-side comparison of IM-Net with SepConv and TOFlow on two in-house FHD test sequences. In Fig. 1 in the main manuscript we showed results for one interpolated frame from each of these sequences, which were originally captured at 30fps. We dropped every other frame to obtain a 15fps sequence. To up-convert them back to 30fps we applied each of the three methods. It is important to note that these two sequences are very challenging: the 'Soccer Bouncing' sequence involves both strong camera motion and complex local motion, and the 'Walking Near Cam' sequence consists of strong object motion.

In the 'Soccer Bouncing' sequence there are noticeable temporal artifacts for the two previous methods. These videos suffer from flicker, wobble and inconsistent object shapes. On the other hand, IM-Net shows no flicker or wobble. Nevertheless, we observe a moderate level of halo and drift of object motion to the background, due to inaccuracies in the estimated IMVF. In the 'Walking Near Cam' sequence, IM-Net obtains consistent object shape and maintains a high quality video throughout, while the other methods fail to preserve the shape of the people.

[1]The avi files were generated with the standard H264 codec and can be played with any popular video player.